SYSTEM AND METHOD FOR CONTEXT-DEPENDENT PROBABILISTIC MODELING OF WORDS AND DOCUMENTS

ABSTRACT OF THE DISCLOSURE

A computer-implemented system and method is disclosed for retrieving documents using context-dependant probabilistic modeling of words and documents. The present invention uses multiple overlapping vectors to represent each document. Each vector is centered on each of the words in the document, and consists of the local environment, i.e., the words that occur close to this word. The vectors are used to build probability models that are used for predictions. In one aspect of the invention a method of context-dependant probabilistic modeling of documents is provided wherein the text of one or more documents are input into the system, each document including human readable words. Context windows are then created around each word in each document. A statistical evaluation of the characteristics of each window is then generated, where the results of the statistical evaluation are not a function of the order of the appearance of words within each window. The statistical evaluation includes the counting of the occurrences of particular words and particular documents and the tabulation of the totals of the counts. The results of the statistical evaluation for each window are then combined. These results are then used for retrieving a document, for extracting features from a document, or for finding a word within a document based on its resulting statistics.